**Title:** Twitter Sentiment Analysis of Changing Trends: Forecasting Sentiment Over Time

**ABSTRACT**

People use social media platforms, such as Twitter, to communicate their thoughts, emotions, and responses to numerous events and issues. The vast amount of data created on Twitter every day provides a wonderful chance for researchers to analyse public mood and identify emerging patterns in real-time. This research focuses on sentiment analysis of Twitter data to investigate how sentiments towards specific subjects or events change over time. In this research we will use a dataset called Project Data that contains data for the year 2009, the analysis in this project will be presented as well as visualized using appropriate figures, tables, captions as well as interactive dashboard for easier interpretation. Eventually, we will forecast the sentiments types for different time periods of one week, one month as well as three months.

**INTRODUCTION**

Social media stages have developed as capable communication channels that encourage the trade of thoughts, suppositions, and feelings on a worldwide scale. Twitter stands out as one of the foremost utilized stages, with millions of clients sharing their contemplations and encounters each day.

The reason of this project is to analyze Twitter’s text sentiments and get it the collective feelings and behaviours of clients inside a particular time period. By analyzing the sentiments communicated in tweets of our collected dataset, we will be able to reveal designs, patterns, and shifts in open conclusion which can be important in making suggestions.

The importance of analyzing Twitter assumption lies in its potential to supply real-time bits of knowledge into open responses and conclusions. Twitter also offers a enormous and different dataset that captures conclusions from a wide run of people over geographies, socioeconomics, and cultures. In addition, assumption investigation on Twitter can also help in making data-driven choices, such as understanding client input, measuring brand discernment, and distinguishing developing patterns. The chosen time period for this project can also be crucial as it will empowers us to recognize patterns and changes into the advancing nature of open states of mind.

**DATA COLLECTION**

**First by definition, Data collection**is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes *(Office of research integrity, 2003)*. In this project we have used the data called ProjectTweet which is in a csv format, and the same data is also available on the internet website [www.kaggle.com](http://www.kaggle.com/). In our data, we have a series of features and that includes Sentiments, IDs, Date, Flag, User and TweetText and this can be seen when the data has been loaded into our Jupyter Notebook, our dataset is imported to our Jupyter Notebook using all relevant libraries before being analysed.



*Figure 1: Data Collection.*

After importing our data, we can see that each row in our dataframe represents a single tweet and our columns has features of Sentiments, IDs, Date, Flag, User and TweetText. This imported csv file will be used for our different analysis including sentiment analysis in order to gain insights into opinions as well as sentiment related to our topic for stated specific period we stated earlier of one week, one month as well as three months.

**TEXT PREPROCESSING**

This is another important task we will look at in our project under data exploration after collecting and importing our data, by definition according to *(Mohanty et al., 2023)* text processing includes the elimination of spaces, line changes and special characters presented in the texts, for ensuring that they did not interfere with the next tasks, text tokenization and analysis of part of speech. Under text processing in our project, in order to prepare the text data for model building, we went through five steps of text processing which involves correcting spellings, removing all links in our text, removing imoji’s, eliminating stops words and lemmatization. Text processing do not only help to get through the steps below, but it also helps in making data more manageable in making further analysis in machine learning algorithm.

**Steps Taken to Clean Data And The Importance:**

1. ***Spelling Corrections.***

Spelling corrections are vital to progress the accuracy and quality of content information. We are able utilize different libraries or calculations to perform spelling rectifications. One common approach is to utilize an outside word reference or dialect show to recommend adjustments for incorrectly spelled words.

1. ***Removing links.***

Links (URLs) in content information are frequently not important for many NLP assignments and can include noise to the examination. You'll be able remove links using customary expressions or string matching techniques.

1. ***Removing imoji’s.***

Emojis are graphical characters that might not include much value to any content investigation, and imoji’s can be expelled using customary expressions or by using specific libraries for imoji’s.

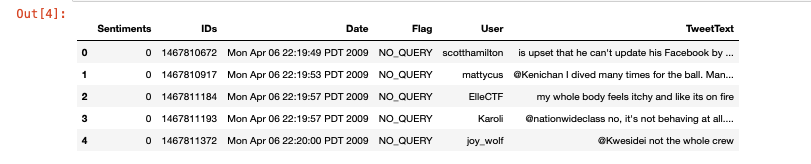
1. *R****emoving Stop Words.***

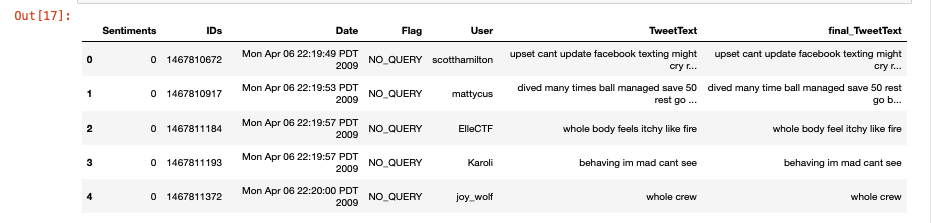
Stopwords are the commonly used words and are removed from the text as they do not add any value to the analysis. These words carry less or no meaning. *i.e.* *i, me, my, myself, we, our, ours, ourselves, you, you're etc*. *(Deepanshi)*

1. ***Lemmatization.***

Lemmatization is the process of gathering word-forms *give, gives, gave, given, giving,* and probably *to give,* will conventionally be lemmatized into the lemma give. Any occurence of any of the six forms will be regarded as an occurence of the lemma.(Sinclair 1991: 173)

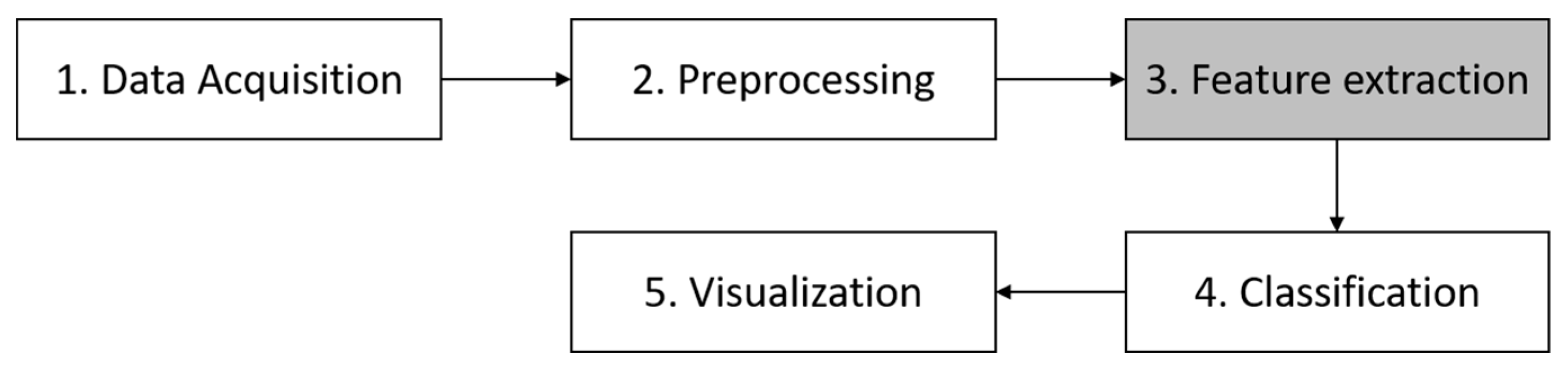
The tables below shows how our data appeared before and after going through our steps of text pre processing. As we can see in our TweetText columns, our text has been cleaned of all unwanted characters, words etc after performing text processing.

 *Table 1: Before Text Processing.*

 *Table 2: After Text Processing.*

**FEATURE EXTRACTION**

This is another segment we looked at in our project, and this involves the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set (uk.mathworks.com, n.d). The figure below shows the strategic plan that involves data acquisition. Preprocessing, feature extraction, classification as well as visualization.



*Figure 2: Research strategic plan.*

The main objective of feature extraction is to represent the data in a more compact and instructive way, capturing the important designs and characteristics of the information whereas lessening its dimensionality.

Within the context of machine learning, the method of feature extraction regularly includes various steps. However, under this chapter of feature extraction we mainly checked number of words, number of stopwords, number of hashtags and number of characters in all our data. First created a copy of data, the data we derived from preprocessing stage, The reason for making this copy of data is that we would like to avoid all modifications or changes made to our data not affecting the original data. From here, we counted and calculate number of stopwords in our dataset, the reason for this count is to analyse any trend in the tweets based on the sentiments. The same thing will also be done on counting and calculating number of words, number of hashtags and number of characters in our dataset, we then eventually use feature extraction function to combine all the functions that consists of all these counts stated to account for a final result that shows a result in reduction of amount of data from our dataset. Feature extraction also recognizes the foremost segregating characteristics in signals, which a machine learning or a profound learning calculation can more effectively expend.

**SENTIMENT ANALYSIS**

Sentiment analysis also called opinion mining is the study of peoples, opinions, sentiments, appraisals, attitudes and emotions towards entities and their attributes expressed in written text. (Bing Liu, 2020). Our main aim of this section of the topic is to be able to determine the ratio of negative to positive interaction about our topic. In sentiment analysis we will be able to analyse our body of Text from our dataset like TweetText or comments to obtain insights from twitter users, and this is also made possible by importing and using all NLTK which will be essential in processing text data that we will use to perform the sentiment analysis on our data.

First we will introduce Sentiment Intensity Analyzer (SIA), the sentiment intensity analyzer is a function that helps in deciding how strong the sentiment is communicated within the context. Sentiment Intensity Analyzer allocates numerical score from -1 to 1, where -1 means a highly negative sentiment, 0 means a neutral sentiment as 1 means a highly positive sentiment.



*Figure 3: Sentiment Analysis.*

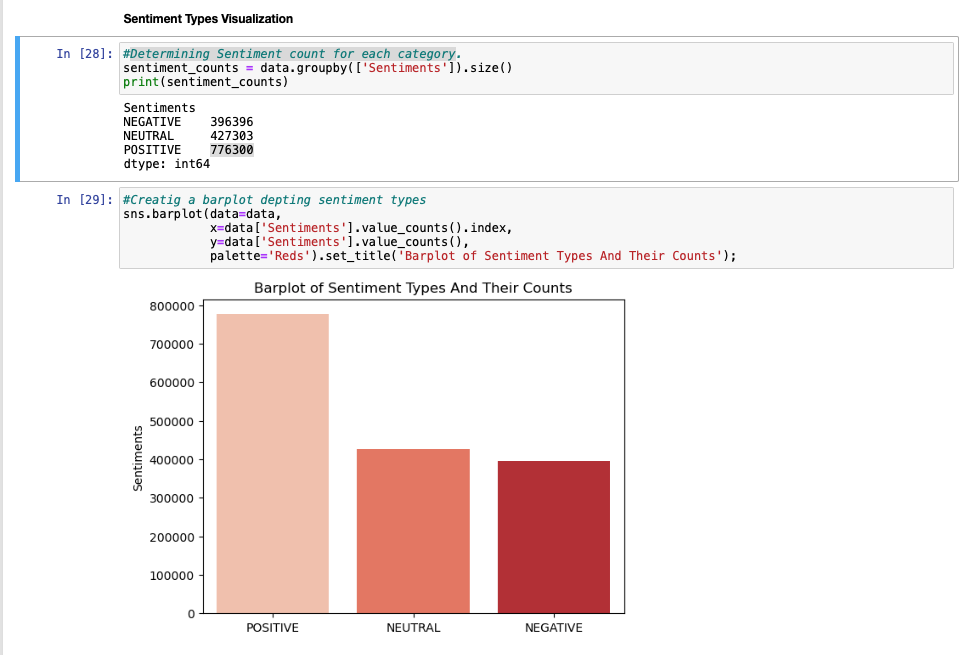
After introducing Sentiment Intensity Analyzer, we will then assign polarity score. Polarity score is a library that returns a compound score, this is also a measure that calculates the total of all normalized lexicon ratings. From assigning polarity score we then assign sentiment types based on what number the polarity takes. In our research we categorized our sentiments into positive, negative and neutral types. This means if the sentiment compound polarity is greater than 0 *(>0)* the polarity will take a positive sentiment type, if the sentiment compound polarity is less than 0 *(<0)* the polarity will take a negative sentiment type and finally if the sentiment compound polarity is equal to 0 *(==0)* the polarity will take a neutral sentiment type.

The T*able 3* below shows the results after assigning polarity score to each tweet as well as assigning sentiment types based on what polarity number takes.

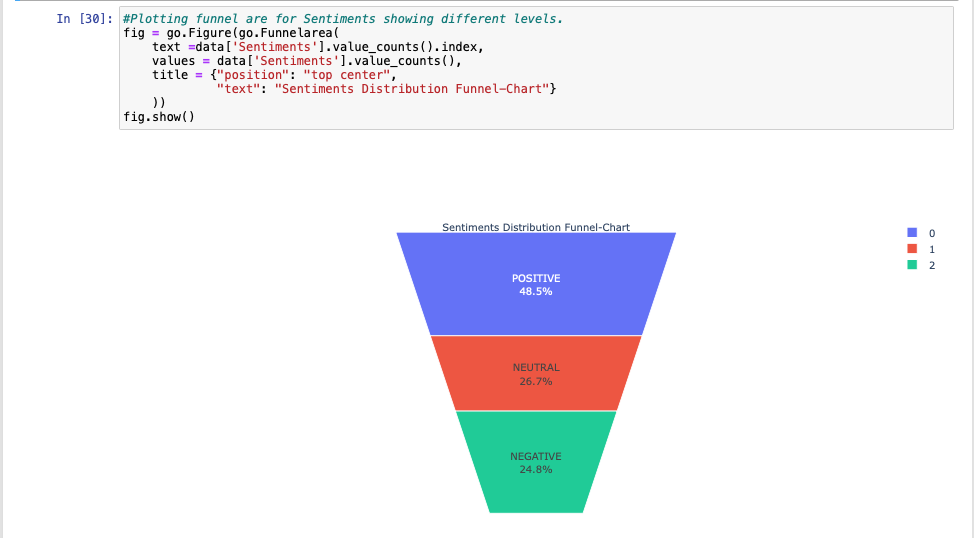
 *Table 3: Sentiments and Polarity Score Assignments.*

**EXPLORATORY DATA ANALYSIS (EDA)**

This is another step we took in our project, Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.([www.ibm.com](http://www.ibm.com/), n.d). Before diving into making our prediction as well as analysis, we went through exploratory data analysis with the goal of gaining more insights as well as understanding of our data. First we did so by determining sentiments count for each category we have in our dataset before our visualization. From the results, we can see that of all sentiments the positive sentiment has higher count than the rest in our dataset with a count of *776300,* the second is neutral that has a count of *427303* and negative sentiments has a count of *396396.* We have visualized well this result using the bar-plot as seen in the *figure 4* below.

*Figure 4: Sentiment types and its Counts.*

As seen from the figure above we can clearly see how the sentiment types have been distributed according to their counts with positive sentiment higher than all, then neutral as second as negative sentiment as the least of them all with less count. Another plot we presented in this project was Sentiment Distribution Funnel Chart, this kind of plot will also help in providing more understanding the balance of positive, neutral as well as negative sentiments in our data. From the next figure *figure 5,* funnel chat we can see how the count has been distributed in percentages. Positive sentiment with already higher count has been presented with a *48.5%* of counts on the first horizontal line while neutral takes *26.7%* on second horizontal line and the rest *24.8%* represents the negative count on our sentiments, for an extra information on this visualization we can also hoover around the visualization area for more visual information.

 *Figure 5: Sentiments type distribution in percentages.*

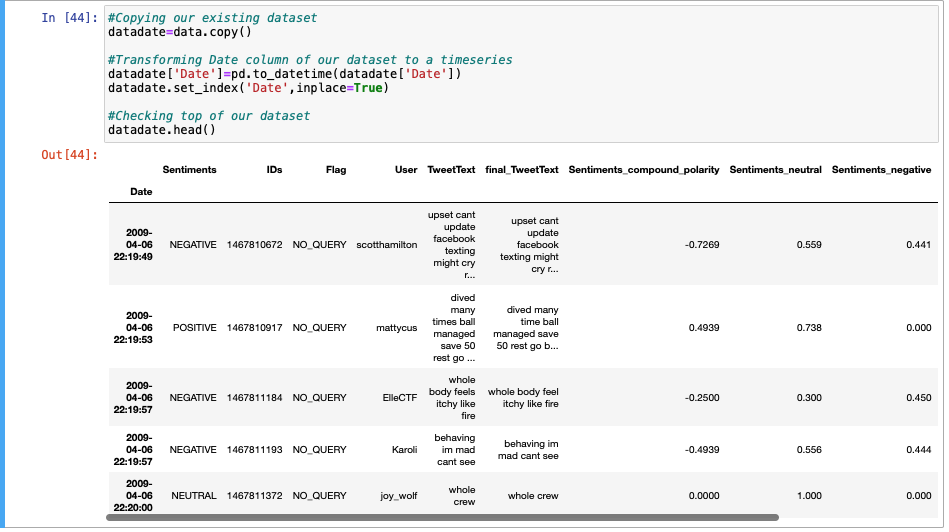
Another tradition step we took in our exploratory data analysis was to retrieve most common words in our dataset, the main reason for this is to see how many words are associated with sentiments. By computing our EDA function as well as visualizing the results we can see that the word *“im”* is the most common word in our dataset with a count of *177.754k* than the rest, and since we selected only top 10 to be visualized, we can also see that the word *“dont”* is the least common word according to our visualization and this has a count of *66.951k*. And for more insight, this can also be presented using a tree map for more information.

We also presented most common words by sentiment types in which *“im”* was also the most common word in most common positive words with a count of *88247,* followed by *“go”* as the least common word of the top 10 with a count of *37398.* And the same result is similar in most common neutral as most common negative words.

We also took upon wordcloud as another form of visualization in this research. Word Cloud is simply a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance *(GeeksforGeeks, 2018)*. We created wordcloud so as to get most sense of the words associated with our sentiment type whether positive, negative as well as neutral sentiment. By looking at our wordcloud visualization results for all sentiments in our Jupyter Notebook, we are able to tell which worlds appear most as positive, negative as well as neutral in our dataset. Wordcloud provide us with a lot of advantages during exploratory data analytics such as helping us gaining more insights in this text in determining the type of sentiments we have in our dataset, they are also powerful in visualizing what internet users i.e twitter in our case thinks/share their opinions about topic.

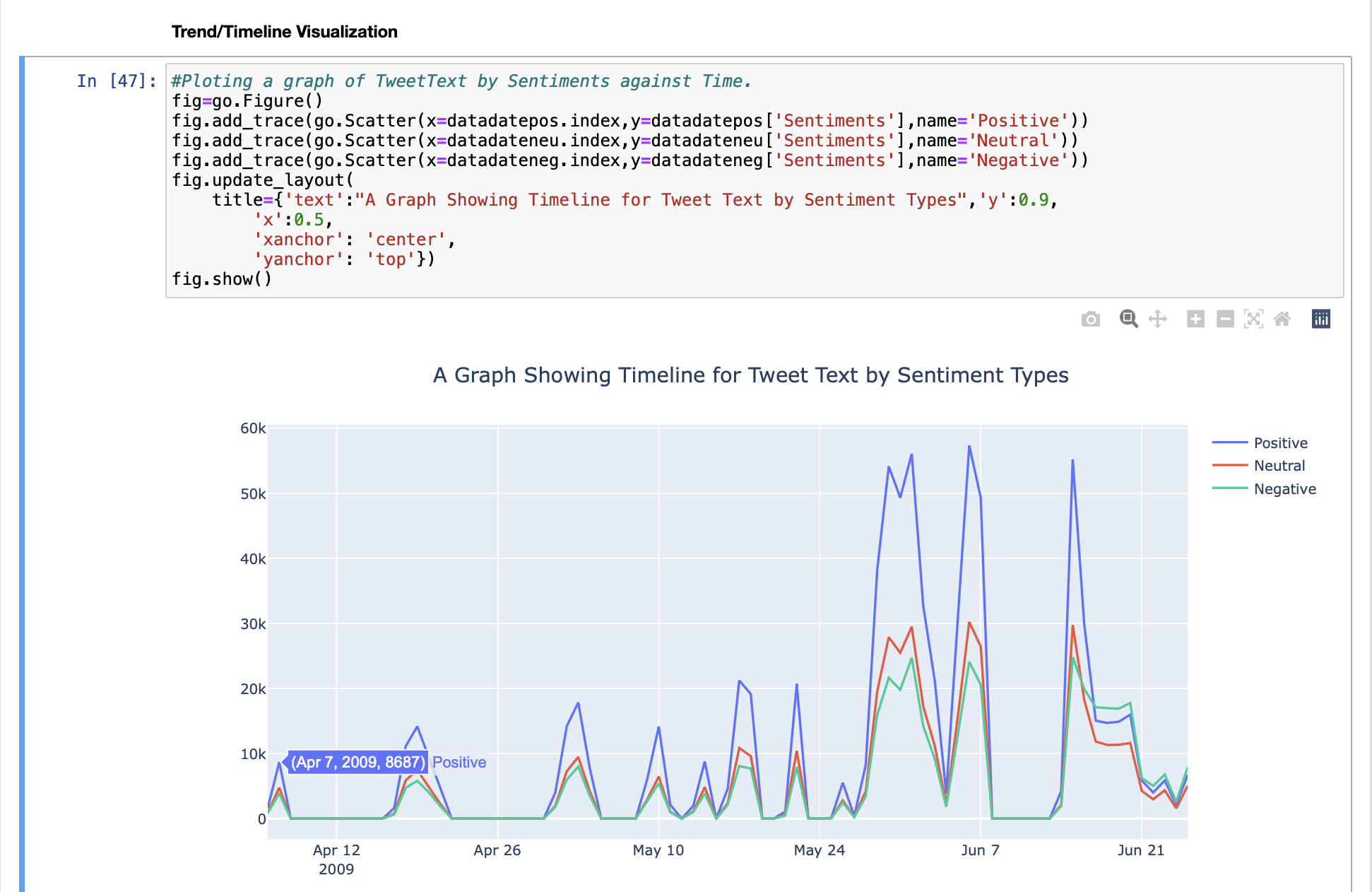
**TIMELINE ANALYSIS**

Twitter trends is the examination that includes observing the popularity and recurrence of particular watchwords, hashtags, or themes on Twitter. Twitter patterns are energetic and can change quickly based on real-time occasions, viral posts even breaking news. Analyzing Twitter patterns can give bits of knowledge into what points are right now well known or what occasions are picking up critical consideration. In thus section we copied our data with the code *datadate.data.copy()* with the same reason of preventing any alteration or modification affecting our original data. The first essential step we took in this chapter before timeline analysis was to make sure the date column we have in our data is being transformed to a time series, in time series analysis time is an important variable of data as it helps us to study the progress of every observation.

 *Table 2: Transforming Date Column to Time series.*

From the above figure, we can see the result after date column conversion to a time series. In this way on our project, we will be able to understand the underlying patterns or trends that changes overtime using data visualization.

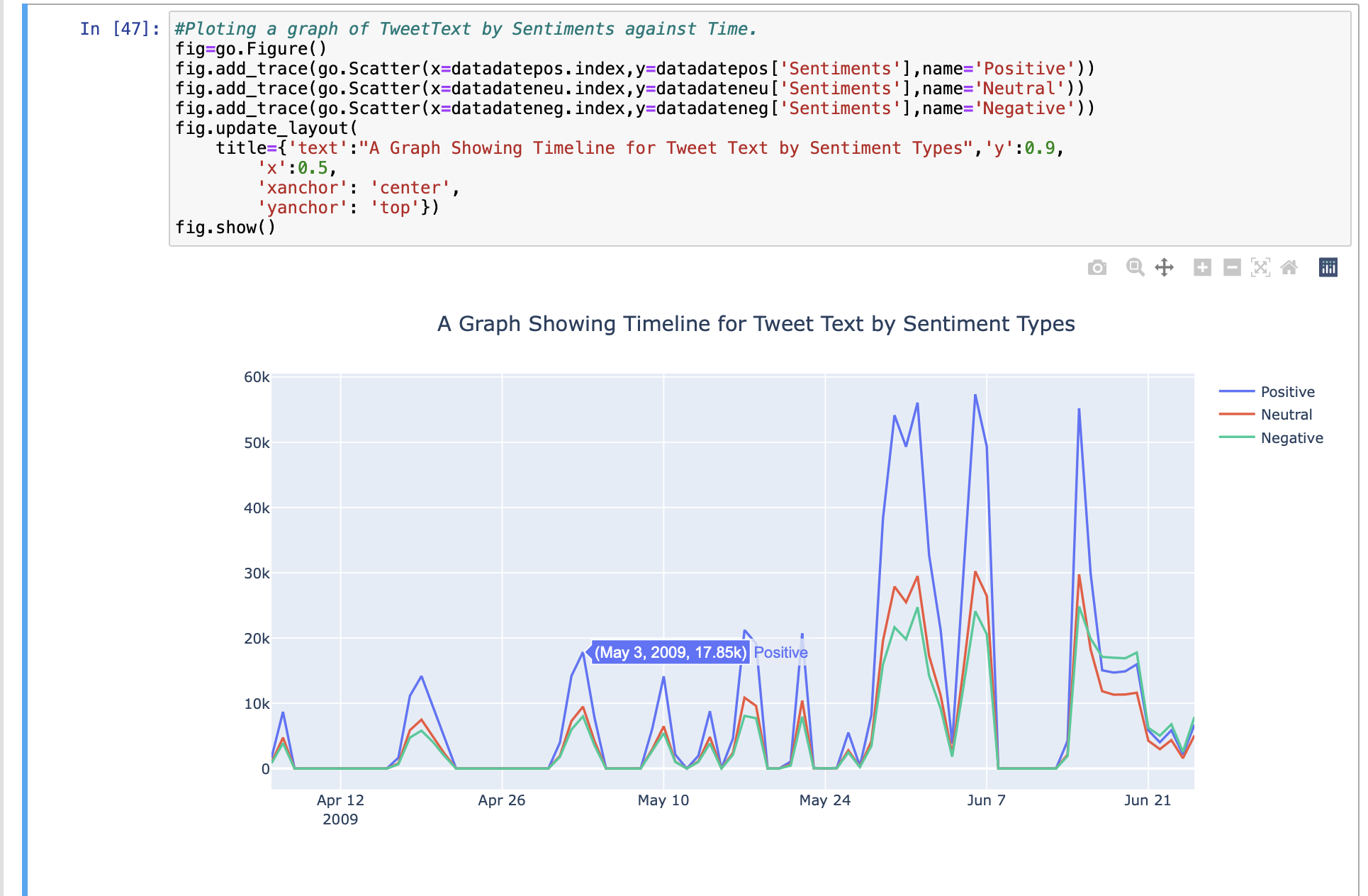
After date conversion to a time series, we also created different datasets based on sentiments time we have. These different datasets will help us to be able to to visualize all types of sentiments *(positive, negative and neutral)* over time, and after this we will resample all newly created data so we can easily be able to to create our trend. The code on the *figure 6* below shows a code we used to plot our graph for tweeter sentiments over time.

 *Figure 6: Sentiments results after 7 days.*

**RESULTS AND DISCUSSION**

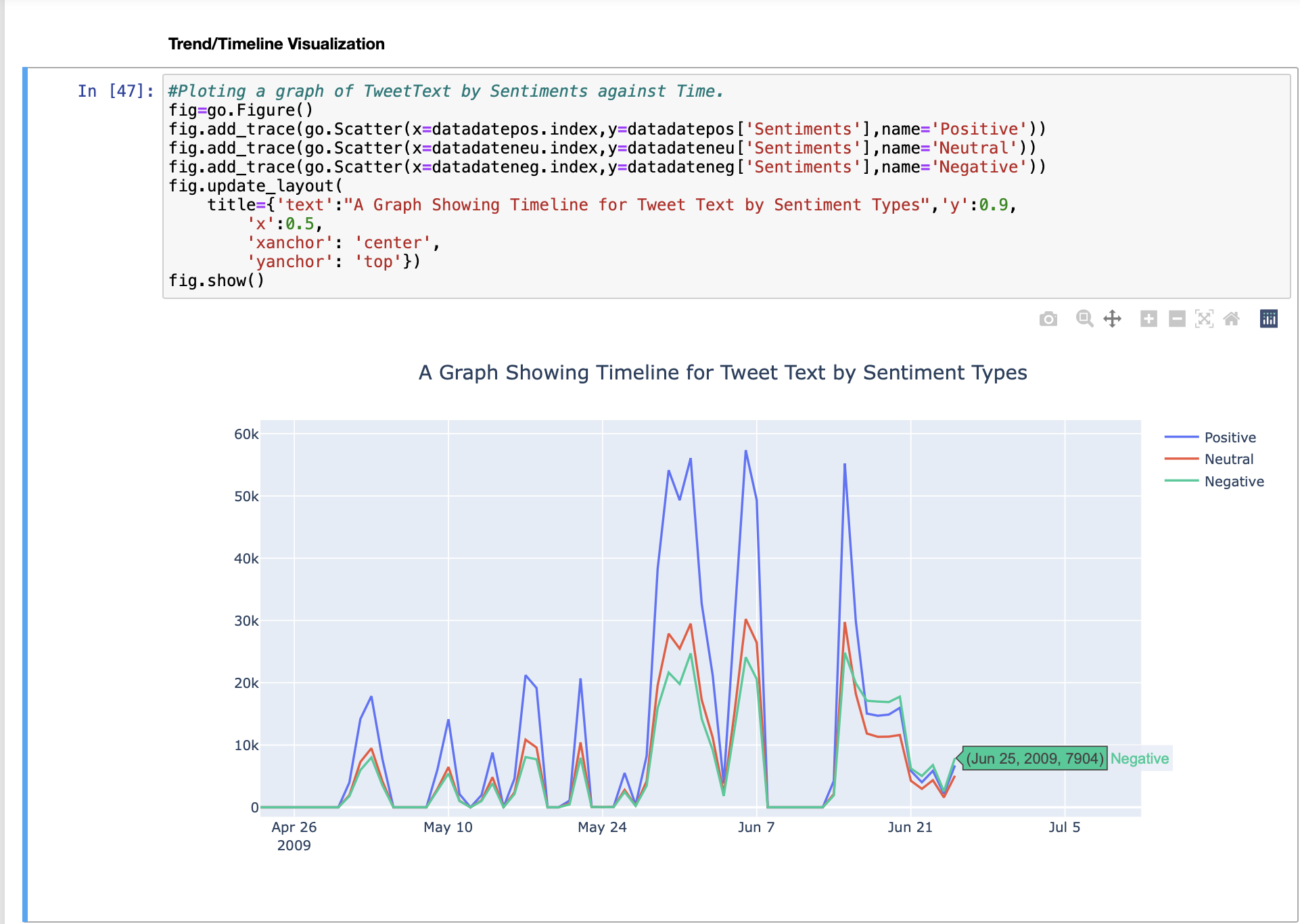
From our results in figure 6 above and on the Jupyter Notebook, we can see that the first week *(After 7 days)* on our plot there was a big incline in positive tweets than both neutral and negative with number of tweets at *8687k* on April 7th 2009, the second highest incline was neutral with *4739k tweets on the same day and lastly on the 7th* daywasnegativewith *3885k* tweets. We can see this all by hoovering the cursor around the our plot in our Jupyter Notebook, we can also choose to only view one or two sentiments only by selecting the ones we would not like to see on our visualization in the legend section. Or legend is defined by three colors, red for neutral, blue for positive as well as green for negative.

Similarly, after *(After 30 days)* we can also see another big incline on positive comments again from *0* to *14.184k* and then to *17.85k* while neutral sentiment this time stares at *7280k* from *0* (After *30 days)* and negative sentiments at *5970k* from *0* after same amount of period. The figure 7 below shows changes in tweets for all different sentiments (Positive, Negative and Neutral) after 30 days.



*Figure 7: Sentiments results after 30 days.*

Lastly, in our time series analysis we also looked at changes in sentiments *after 90 days*, and this time we can see that after there were series of up and downs of tweets during the period until 24th June before it begin to start going up again. After 25th June we can now see an increase in all sentiments type, but this time negative sentiments appears to be inclining more than the rest with *7904k* tweets while positive sentiment has tweets has *6723k* after 25th June and neutral has *5067k* tweets which appears to be lower than the rest.

 *Figure 8: Sentiments results after 3 months.*

**CONCLUSION**

The conclusion of the analysis in this project summarizes the forecast that comes about and highlights the key experiences and potential patterns anticipated on Twitter's timeline. It may moreover offer recommendations for businesses or people looking to use the anticipated patterns for their advantage.

From our results, we can see that the overall our tweets produced more positive tweets of our topic than the negative as well as the neutral. Generally, anticipating long-standing time of Twitter timelines requires steady checking and examination of current patterns and stage advancements. Modern highlights, changes in client behavior, and external events can all impact the flow of Twitter timelines and ought to be taken under consideration in any determining investigation.

**REFERENCES**

1. Office of research integrity (2003). *Data Collection*. [online] Hhs.gov. Available at:

<https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html>.

1. www.telusinternational.com. (n.d.).*Five pitfalls to avoid before you start collecting data*. [online] Available at:

<https://www.telusinternational.com/insights/ai-data/article/five-pitfalls-to-avoid-before-you-start-collecting-data>.

1. Sachi Nandan Mohanty, Vicente Garcia Diaz and Satish, E. (2023). I*ntelligent Systems and Machine Learning*. Springer Nature.
2. Monkeylearn (2018). *Sentiment Analysis: Nearly Everything You Need to Know | MonkeyLearn*. [online] MonkeyLearn. Available at:

<https://monkeylearn.com/sentiment-analysis/>.

1. Deepanshi (2021). T*ext Preprocessing NLP | Text Preprocessing in NLP with Python codes*. [online] Analytics Vidhya. Available at:

<https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/>.

1. Flowerdew, L. (2008). Corpus-based analyses of the problem-solution pattern: a phraseological approach. Amsterdam ; Philadelphia: John Benjamins Pub.
2. uk.mathworks.com. (n.d.). Feature Extraction. [online] Available at:

<https://uk.mathworks.com/discovery/feature-extraction.html>.

1. Bing Liu (2020). Sentiment analysis : mining opinions, sentiments, and emotions. Cambridge ; New York: Cambridge University Press.
2. www.ibm.com. (n.d.). What is Exploratory Data Analysis? | IBM. [online] Available at:
3. [https://www.ibm.com/topics/exploratory-data-analysis#:~:text=Exploratory%20data%20analysis%20(EDA)%20is](https://www.ibm.com/topics/exploratory-data-analysis" \l ":~:text=Exploratory data analysis (EDA) is).
4. GeeksforGeeks. (2018). Generating Word Cloud in Python. [online] Available at:

<https://www.geeksforgeeks.org/generating-word-cloud-python/>.

Version Control Address: [https://github.com](https://github.com/MphatsoChintedza)